Informal reflections on audit issues surrounding “Cash on Delivery aid” in the education sector

Luis Crouch
RTI International

February 2008

Introduction and background

The Center for Global Development has proposed a “Cash on Delivery aid” (PBA) approach, whereby, in a nutshell, countries could be rewarded for making progress on some internationally-agreed indicators, and on the basis of a contract-like mechanism (or an actual contract). The advantages of such a proposed approach are many, but a few of them would be: a reduction in transactions costs and costs of donor bureaucracy, a clearer signalling from the international community to the developing countries as to what the international community considers important, and a way to attach a monetary value to what might otherwise be seen as a merely ethical desideratum. The latter point is not merely a matter of creative “incentives” in some simple-minded way, but is also a way of creating a mechanism that allows interest groups who advocate for development, within their countries, but often only from a moralist or ethical point of view, to point out to their governments that “doing the right thing” carries immediate economic rewards for the country. This makes the incentives line up with the political economy of pressure groups, in a way that perhaps will more clearly favor development goals.

The education sector has been proposed as a possible “test case” for this approach, both as a thought experiment at first, and possibly as a real experiment if the thought experiment gains traction. In this sector, the basis of the contract would be some commitment to increase primary completion rates (and possibly secondary school entrance rates), and to ensure that this be done subject to some form of quality control via student assessments.

The Center for Global Development proposed to Luis Crouch, of RTI International, a consultancy to explore the audit or audit-like aspects required to ensure contractual compliance around these issues. Lacking time for a proper assessment, he suggested that he and some colleagues would simply lay out some initial ideas, that could have some merit on their own, or that could be provided as input to someone else who could provide a more complete consultancy.

In all of this we think it is wise to simply come out and say that one is really talking about audits as you have done. However, as will soon be seen, we believe that some mix of information produced by firm-based audits, or audits as traditionally-understood, plus the social pressure brought to bear by “social audits” and the disclosure of information, is

---

probably a better idea. This concept is implicit, but implicit only (it seems to us) in the PBA concept note.

**Likely “gaming” issues that audits should try to prevent**

To try to make the audit need clearer, we will try to lay out the issue in a series of cascading indicator gaming strategies. We will suppose that every attempt at gaming the indicator requires an audit strategy, which then leads to another level of gaming, etc. We will leave the likely audit strategy for each level of the game for a later section, and concentrate for now on the game itself.

The most obvious gaming is of course simply to misrepresent the number of completers: to try, in some direct and simple-minded way, to get paid for what has not been done. This could be as crude as to create “ghost” completers. There are ghost teachers and even whole ghost schools in developing countries, so where there is no real audit, creating ghost learners is a piece of cake. This then leads to a need to audit the completion numbers.

If the numbers are audited, the reaction then will be to try to “cheapen” what is meant by completion. The simplest way to do this is by lowering quality overall.

This can be done, in the first instance, by lowering the barriers to progress between grades, for example, which tend to cause dropping out. In fact, in many countries social or automatic promotion policies exist already, but frequently parents and teachers refuse to practice it, since they believe (rightly or wrongly) that repetition helps and that children should not advance from, say, first to second grade if they cannot read at least their letters. The government might then simply try harder to enforce the social promotion strategy, for example. Note that this has two effects: a once-off effect of putting more children through to completion immediately, as less children repeat, and a permanent effect produced by having less dropouts, since it is known that repetition tends to cause dropping out.² The system could also react by having more children skip grades. This will increase the number of completers, but only as a once-off effect. Any of these strategies will have the effect of either permanently increasing the number of completers (to use a hypothetical example) for some reason no 6th graders repeat, there is no dropping out between grades 5 and 6, all 5th graders repeat exactly once, and the system has no population growth. In that case in any given year enrollment in grade 5 will be exactly twice as high as in grade 6 (or, if there was steady state population growth, enrollment in grade 5 would tend to always be higher than that in grade 6, by a factor of \((1 + r) + (1 + r)^2\). If one then were to “force” the repetition rate in grade 5 to zero, enrollment in grade 6, and hence completion, would literally and exactly double the next year. But this could happen only once, evidently, and enrolment in grade 6 would go back down to its constant (or steady state growth) level the moment the bulge passes through. If the lowering of repetition were to be across all grades and less dramatic, the once-off effect would be more diffuse through time, but would still be only a once-off effect that would “wash through” the system, unless the repetition causes dropping out (which, as noted, it does to some degree). The effect of encouraging skipping is entirely like the effect of reducing repetition and is thus also only once-off. All this assumes the system is not supply-constrained. If the bulk created by repetition is preventing entry, i.e., if repetition is constraining supply in a permanent way, then reducing repetition would permanently increase completion, even if it did not reduce dropping out.

² To see why there is a once-off effect, think of a system where the completion level is grade 6, and where (to use a hypothetical example) for some reason no 6th graders repeat, there is no dropping out between grades 5 and 6, all 5th graders repeat exactly once, and the system has no population growth. In that case in any given year enrollment in grade 5 will be exactly twice as high as in grade 6 (or, if there was steady state population growth, enrollment in grade 5 would tend to always be higher than that in grade 6, by a factor of \((1 + r) + (1 + r)^2\). If one then were to “force” the repetition rate in grade 5 to zero, enrollment in grade 6, and hence completion, would literally and exactly double the next year. But this could happen only once, evidently, and enrolment in grade 6 would go back down to its constant (or steady state growth) level the moment the bulge passes through. If the lowering of repetition were to be across all grades and less dramatic, the once-off effect would be more diffuse through time, but would still be only a once-off effect that would “wash through” the system, unless the repetition causes dropping out (which, as noted, it does to some degree). The effect of encouraging skipping is entirely like the effect of reducing repetition and is thus also only once-off. All this assumes the system is not supply-constrained. If the bulk created by repetition is preventing entry, i.e., if repetition is constraining supply in a permanent way, then reducing repetition would permanently increase completion, even if it did not reduce dropping out.
completers, or at least having the once-off effect of bringing forward in time the numbers of completers, which one would expect, if governments discount the future. Since there is some evidence that repetition does not really improve cognitive performance, reducing repetition, as a way of lowering the cost of completion, might not be a bad thing. It seems unlikely that a wholesale increase in grade skipping could leave quality unaltered.

Another approach is to have even less qualifications for what it might take to complete. Thus, a child who is in school for only 20% of the last grade of primary, or has attended only 30% each year on average, could be passed along and could “complete.”

Other and perhaps more obvious ways to “cheapen” what is meant by completion of course is just to massively open up bad schools with poorly trained teachers, become even more lax with time-on-task and attendance, reduce the amount of inspections and supervisions per school, provide less books, etc., and make the children just pass through the system regardless of how little they learn. There would presumably be an impact on quality.

One could try to prevent this by having all kinds of input quality and quantity indicators and ensuring that they stay constant per child, but this would be very complex, and in many ways sends the wrong message, because it tends to imply that one knows the input-output relationship (or the “production function”) which one really does not, and anyway it defeats the whole purpose of the exercise, because it is hardly “hands off.”

So, one tries to prevent the gaming not by focussing on inputs, but on outcomes, by requiring some sort of primary school exit exams, which in fact already exist in many countries.

The PBA concept note proposes simply that students be tested, leaving the issue of an average grade up in the air. We assume this is done to avoid the next “stages” in a gaming process, that we discuss below—“to avoid creating incentives to manipulate results” is the language used in the PBA concept note, but the problem is a bit more complex than that. Actually, in the version of the concept note we have, the statement of this issue is a bit vague, saying that payments would be done “without linking payments to test scores.” This could be interpreted to mean that higher average scores would not result in higher rewards (and lower scores would result in lower payments) but could also be interpreted to mean that countries would get the payments even if average scores actually go down. This needs to be made clearer, in our view, in the concept note.

3 This is a more important point than at first appears. Our assertion is that when economists and other donor “interveners” take a social-planner approach, instead of an outcomes-denominated “social contract” approach, they implicitly or explicitly assume they have rather good knowledge on what the production function really is, or that the more subtle elements in the production function, such as the kinds of effort that are really usually only created by incentives (whether monetary, social-esteem incentives, or professional pride as an incentive), can be “supplied” or thought of as an “input.” We would assert that either we do not know what the production function is, with more than about 40% uncertainty, or that some of its most important elements are not mere “inputs” that can be supplied, but have more to do with effort and creativity, which cannot be supplied but can only be stimulated through a system of social incentives and controls.
Presumably, then, what the PBA intends is that one would rely on general social pressure, endogenous and internal to the country, to ensure that the average grade in the test stays up, and the government is not “cheating” by lowering cognitive demands on the test. This is probably healthy, but the proponents of the PBA approach, we believe, do have to ask themselves what would happen if, say, the average mark were to go down by 10% at the same time as the number of completers went up by 10%. Would that really be acceptable?

It should also be noted that many countries do not really use a notion of an average grade on exams, but use the notion of the pass rate. This is relatively easy to manipulate. The PBA concept note uses the notion of an average grade; it should stick (consciously) to that notion, and discourage ideas of focusing on a pass rate.

If the results or the exams are merely to be published, but there is no reward tied to the results, I would still suggest that what should be published is something more like the average mark, not the percent passing. Or both. But one really wants to prevent the gaming around the percent-passing that I have noted below.

Since the PBA concept note does not require average grades to say up, the following few paragraphs are perhaps not relevant. We include them in case the authors change their minds.

Suppose one decides that social pressure or social audit, around merely publishing the results of the exam, is not enough.

Then there are two further requirements one could impose. First, to require that the pass rate not go down. This would be the most typical barrier a government would agree to impose, because reporting and focusing on pass rates is the most common practice. This is then gamed by improving quality very selectively in those schools whose average grade is right below what is needed to score a pass (as an economic proxy for targeting those children whose exact grade is close to what is needed for a pass). One can produce a significant gain in the percentages passing (or, rather, prevent a drop in the percentages passing) while not affecting the average grade at all, or even while the average grade goes down.

4 Because most countries use exams as a filtering system, they focus on a pass rate, with a notion of a minimum grade needed to pass. The average grade needed to pass (or the cognitive demand of the exam) is sometimes actually manipulated to control the numbers passing, because the numbers of seats available at the next level are fixed. Thus, the pass rate and the numbers sitting for the exam sometimes have an inverse relationship with each other. That is, exams are graded on a “curve” and the curve is adjusted to what is needed.

5 Suppose the minimum pass grade is 60 points. Suppose 1/3 of the schools average 30. Suppose 1/3 average 55. Suppose 1/3 average 80. The pass rate is then 33.3%, and the average grade is 55. The system could shift its best resources to the schools whose average grade is 55, such as to cause them to “tip” right over 60, say to 65. There would be no point in helping the schools whose average grade is 30, as it would be too difficult to get them to 60. One could then have the pass rate go up from 33.3% to 66.7%, thus doubling, while the average mark has shifted from 55 to 58, thus increasing by only 6%. Or, if the withdrawal of resources from the better-off schools causes an exactly compensating drop in the average mark, the pass rate would go up 6.6% to 68%, while the average grade went down by 7%.
Second, one could stipulate that the average grade (or average mark as they’d say in other countries) not go down.

There could be one initial reaction to the average grade indicator, namely to focus attention on children whose average grades are easiest to increase, or keep from going down, as enrollment increases. If there are diminishing returns to inputs, and inputs were the main cause of performance, this might not be a bad thing as it would tend to increase equality of outcomes. But this is a fairly sophisticated reaction and is unlikely. More worryingly, it may be that some the easier-to-increase children might be those who already do well, so this might increase inequality. Again, though, this reaction also is probably too sophisticated and is an unlikely gaming strategy.

The most likely common gaming strategy to a requirement that the average grade not go down would be to lower the cognitive demand of the exam. One then has to make sure that the cognitive demand is constant across time by psychometrically auditing the exam. (On the government’s side, if this really were a game, they would have to guard that the donor is not trying to get out of having to pay by making the exam more difficult. Thus, some serious psychometric “equating” work will have to be done.) This will require a level of audit or technical assistance that now seems to be departing from a “hands off” approach, but is certainly, still, much more “hands off” and laudable than the current focus on enrollment and inputs.

Then come several reactions.

One, to cheat outright by leaking the test and teaching to the test in a very direct way, we assume is relatively easy to guard against, at least in theory, by working on exam security. Yet, as experience in many countries suggests, this is not at all easy. And if the interested party in leaking, in some sense, the exam, is the system itself, then this raises the level of audit requirements to a very high point.

A second is to “teach to the test.” This might not be a bad thing. If children are not learning much at all, and if the test is rather good, then the narrowing of the curriculum (the taught curriculum) caused by teaching to the test would likely have much higher benefits than costs. It is also likely that with social pressure, and with awareness of the problem, society would over time come to control, to some degree, the “teaching to the test” phenomenon. But this would indeed require a high quality test and fairly extensive test, so that the teaching to the test does indeed have a good effect.

grade in those, then the average grade in the system as a whole stays constant, while the pass rate doubles. One can think of the same “game” for preventing drops in pass rates rather than increasing them. In reality the situation might not be as dire as this, since the reaction of schools’ grades to quality of resources (or amount of resources) probably has diminishing returns at the margin. Withdrawing some of the good inspectors, principals, and teachers, say, from schools that are already doing well, and shifting them to those right under the passing grade, is unlikely to reduce the “good” schools’ average grades by as much as it is likely to increase the average grades in the worse-off schools, but our general point holds.
Other issues related to statistical nature of the audits

Having discussed the various “gaming” issues that might arise, one needs to consider some of the “statistical” aspects related to the indicators and/or to the audit process.

5-year lag on the completion. Was this on purpose? This will create a bump the first year, almost certainly, as the completers will almost certainly be higher than you’d get by applying past completion rates to the current population. If on purpose, fine. If inadvertent, you might want to think about it.

Completion projections. The PBA seems to implicitly assume that the number of completions has a pretty well-defined trend. But what if not? What if there is a fairly broad prediction confidence interval? Would the number of projected completions that needs to be “beaten” be the lower bound on a prediction confidence interval or the central point of the projection? One would think the number that needs to be beaten is the central point of the projection, because the lower bound of the prediction interval would be lower than the point estimate of the last actual year. But, again, that assumes that there is a very simple extrapolation without much of an inherent error because the numbers tend to be have some randomness. Examination of past data suggest that completion numbers are fairly “jumpy” but this could be because countries simply have not minded the issue very much.

One reason for a fairly broad confidence interval might have to do with the population projection that is used. These are, after all, just projections. So you have error in both the numerator and the denominator. You could get rid of some error by just projecting a moving average of the enrolment numbers themselves, or something like that. It may be worth thinking about.

Is this one reason to allow some discrepancy or some “fogginess” in the meeting of the targets?

Furthermore, with regard to the projections, would the methodology used to predict be exactly the same for all countries? (Why a 5-year lag? Why not 6 or 4?) Who will decide on that methodology? What if some methodologies would, ex-ante, be seen to favor some countries? That is, there may well be jockeying, in the choice of methodology, because some methodologies will produce a less steep slope than others, and this will favor more countries. And, what if the prediction confidence intervals for some countries are broader than for others, because the historical data are more random? What does that say regardless of whether the level that must be beaten is the central point of the prediction, or the bottom of its confidence interval? Will that create a perceived unfairness in treatment between countries?

The contract should specify that the sample of students to be tested needs to be random. This is probably implicit, but needs to be explicit.
In either measuring completion, or in the testing, a few additional audit considerations would need to be taken into account:

1. The sample audits, particularly on the testing side, would have to vary year by year, otherwise the governments will of course place selective attention on the “panel” that is being followed year on year, making the exercise worthless. It would take a very short time for the government to discover how easy it is to do this.

2. In addition, the new sample to be used each year cannot be pre-announced, or else, again, there will be selective attention to the schools that will be part of the test. The sample has to be determined only at the end of the school year.

In deciding what range of deviations to allow in either the completion rate or the testing rates (and testing results, were one to go back to that idea), there are two quite separate issues to consider, and the current concept note seems to conflate them. (Actually, there are three. The prediction error is another one, already covered above. It might be wise to think of all three at the same time.) There is the issue of the statistical significance. If a sample finds that the completions are smaller than the completions claimed by the government, what is the margin of sampling error allowed, vs the margin of substantive difference allowed? It seems to us that in the current note the notion of some sort of substantive forgiveness is being indulged in because of the realization that samples of reasonable cost will tend to have some sampling error. But is this right? The reason this is relevant is that ultimately the sampling error can be brought down to an essentially arbitrarily low level by increasing the sample size, whereas the issue of substantive deviation allowed is an a priori policy decision—and why should any substantive deviation be allowed (unless it is because of the issue of a prediction interval)? To allow for any substantive forgiveness sends the wrong message, so the point has to be made very clear, if any forgiveness is to be allowed, that this is due to sampling error (or prediction error, or indeed due to other problems such as a recession). In that case, the implication is that one would have to insist on a very large sample, and tolerate a very small deviation, no? Or, if one knows that very large samples are simply unlikely, then the solution would be to pre-decide on the level of substantive difference Y allowed, and then design the sample size so as to be able to make X% sure that such a difference would not turn up as a sampling artifact. It seems important to us to agree, or determine ahead of time, on both the Y and the X%. The two issues appear to use to be somewhat conflated in the concept note, but that is perhaps a misreading on our part.

The concept note says that data on completion and testing need to be published down to the smallest reasonable jurisdiction. (We hope this was not an idea that we submitted in a previous set of comments!) This might increase the sample size unreasonably, particularly if one needs to drive down the sampling error to a truly minimum level.

And in any case completion is a hard-to-define issue at the very local level, as kids actually move around to schools with more grades, and so on. Thus, some schools or districts may have more than 100% completion, others much less, even though the kids did not drop out—they just went to a different district or school. This is another way of
saying that schools don’t have totally simple and unique catchment areas, nor do districts. Once you get to the province level, maybe. Maybe. I can think of provinces that encompass, say, the capital area, where the completion rate will be higher than 100% in that area, and lower than 100% in the areas around it, but 100% if you take BOTH into account. Not a trivial hassle.

Audit implementation

We do not see any alternative but to use independent firms to do these audits. They need not be international firms, of course, and probably should not be. But they could be supervised by an international panel or an international firm. However, contracting a firm to gather all the data, rather than to do a random audit, would seem extravagantly expensive, and it is difficult to see exactly what one gains. Such a firm would be corruptible—perhaps a little less than government officials themselves, but corruptible anyway. And, the more they have to do, the harder it would be to detect the corruption. A firm charged with auditing the government introduces a secondary actor, is cheaper, and the secondary actor can in turn be overseen more easily because, given that they are doing only a random audit, they themselves would be easier to audit. Thus, to us the conclusion is pretty clear that the best approach is to appoint firms to do random audits, but not to do the actual process of measurement. We would suggest that these firms themselves be audited, at least with regard to the quality of their processes, if not actually their results.

On the notion of doing a random sample of households as such. This is much harder than it sounds. You have to realize that you have to target households that have kids who are completers in them, or who have kids who should have completed. So you need a survey that is highly selective on age. There is no sampling frame you can use for this, unless perhaps you can use the census? And this makes cluster-based sampling difficult, even if you have the sampling frame. So, you’d have to somehow get to the area, and THEN enumerate the households, and find out whether they have kids in the right age group, so you can create your sampling cluster.

I don’t think you can do a HH survey and just ask “what was the highest grade completed” and look for movement in that, because that will apply to the entire population, not the exact age cohorts that are in the denominator of the completion rate (it is just a SINGLE age cohort!!). This is a tough technical issue that whomever you appoint would need to think about, if you approach this from the HH survey point of view. But you should be aware of it too, because it will hugely increase the expense.

There seems to us to be little advantage in comparing school records to households on a case by case basis, even for a random sample, and it would be a great deal more expensive that simply random-sampling households and comparing this to the administrative records on aggregate. Perhaps there would be a developmental advantage to doing some comparisons between schools and households, in that the reasons for divergences between schools records and household perceptions could be determined,
and thus school record-keeping would be improved. But this is mission-creep, and tends to betray the “hands-off” nature of the approach.

In any case, what does it mean to complete? That is, what is to be gained by following the children home? To see whether they have in fact “completed” or were tested if the school (government) claims they completed or were tested? This might have some merit, but it places a burden on finding out “who is right” since the notion of completing is not something self-evident that a survey enumerator can detect.

If the government cheapens the meaning of completion by, say, allowing someone who attended only X% (where X is low) of the time to still “complete,” the HH might disagree. It happens fairly frequently (e.g., in the case of automatic promotion) that HHs disagree with schools, and ask that their kids repeat a grade anyway. Might HHs sometimes disagree with schools if they feel the child has not attended enough to be declared a completer? And in that case who is right?

Thus, this whole effort might call for countries to formalize the issuance of certifications of completion or certifications of testing, for two reasons: first, so that information is sent to the household that the child really and truly has completed (so that if you go check with the HH, the HH can’t simply be wrong due to “lack of information”), and, second, so that the system has to more fully specify what it means to complete—so you get a better definition of the contractual commodity, the commodity around which the contract is built.

This is not the same thing as “paying for vaccinations!”

Note that if the HH claims the child did not complete, but the school says they did, and even if you have a certification process, you still have the possibility that the school will simply claim the HH or the child lost the certificate.

Trying to conclude

I think I’ve made my point that “completion” is not as simple as it sounds. There are options for audit, but they are also not simple, and the researcher you appoint to help you think this through will need to do some really hard thinking about how this can be approached—thinking for which I have not had time. I have tried to do you a service by alerting you to the complexities involved in doing a real audit of these things.

Ward H made the point that if completion is itself such a fuzzy metric, then while the contract may insist on completion, perhaps we should not become too obsessed with it, and put more attention on the testing. That is an excellent point. If both suffer from serious measurement problems, why not focus more, then, on the metric that is ultimately more important?

It could also be that getting too carried away with the audit idea might simply unearth too many worms about how countries are currently manipulating the notion of completion.
This could create a real problem—you might uncover that the countries are actually achieving less completion currently than one thought they were, at least if you hold to a minimum standard of what it means to complete (e.g., a certain standard of attendance), not to mention learning.

So it may be best not to touch this, and just audit to prevent the grossest levels of misrepresentation, such as creating ghost kids. So, all you audit on the completion side is that no one has actually purely invented numbers in the sense of literally invented completers that do not even exist. Checking whether they actually and truly “completed,” and did this in some meaningful way, might just be undoable with any reasonable level of accuracy, and without creating more problems than you solve.

But if we are requiring a test, then perhaps we simply don’t have to worry so much about the completion issue. The testing itself will tell us to what degree completion is becoming cheapened. This does take us back, though, to the huge importance of the test, and ensuring its quality, and comparability over time, in a psychometric sense, and in other senses meant to stop cheating (as mentioned above). Again, for example, you would want to make sure that you random test at unannounced schools. This means that you can’t do longitudinal studies of the effect of the intervention on schools, unless you keep ONE panel for the sake of understanding the dynamics, and one random group with entry and exit for audit or “true” impact evaluation purposes.

So I’d reinforce the notion that in spite of all the possible difficulties, ensuring a high quality test might be in the end easier than getting strict on what it means to “complete.”

I think Hewlett was suggesting that in addition to “simply testing and publishing” one might go back to the notion of tying the incentives to actual performance on the test. I still sort of thing that this might raise the stakes too much, and lead to cheating, and also that one does not know what to expect. If more rural kids are coming into the system, you might actually expect the grades to go down, on average. So this is a really tough problem, and I guess I lean in the direction of not tying the funds to performance on the test, just on having the test (as is the case in the concept note).

And all of this, of course, points to the issue I kept referring to. You just can’t really write a complete contract. The attempt to do so just leads you to trying to specify a contract for how to run an education system properly, so you actually lose the hands-off aspect. And this is related to the fact that the result we are looking for is not an easily specifiable commodity, such as giving a kid a shot in the arm, or buying Red Wheat #2 with at most 10% moisture and 5% foreign matter. What do firms that cannot specify fully a contract? They don’t pay on a piece rate basis, or they do not rely on agents but on employees who are part of the culture of the firm. In that sense, ultimately what will make this work is the pressure and the fillip all this gives to the local forces that will hopefully be vigilant about these things. The contract gets fully specified, in a sense, ONLY through the fuzzy but more complete watchfulness of NGOs and local advocates who will now have sharper, outcomes-oriented reasons to keep their government under
scrutiny, to make sure governments are not over-cheapening what it means to complete, etc.

Couple of last points

1. As we all saw, and even I tended to fall victim to it, the temptation is great to try to fix all problems with this new mechanism. There is a lesson here, and perhaps you might want to commission someone to think about it. Why is it that every simple idea that some creative folks such as yourselves come up with tends to end up trying to solve every problem with the same instrument? Why do these things ALWAYS tend to do back in the direction of hands-on and input-specification? I would suggest that there are two reasons. First, there are some intractable problems, as I’ve noted; you really can’t write a complete contract. So there is a pretty natural tendency to try to make the contract complete. Second, though, I do think there are political economic reasons on the donor side to make things complex. Every transactions cost feeds an internal lobby within the donor community, and every complexity creates a need for one more bureaucrat. You might ask someone to think about and write a note on why this happens, so you can try to keep it from happening, and explain to people why you are so militant about keeping it outcomes-denominated, even if it does mean that the contract is incomplete… That it does assume that other projects are there, and that civil society is there, to round out the contract.

2. I think you are having a lot of trouble explaining the fact that the money is not earmarked but conditional. (Conditional on progress.) And that therefore you do not have to have vouchers and receipts to prove you spent it on education, or, worse yet, the marginal completers! True, you do want to have sufficient fiduciary control to prove that this specific money did not end up being corrupted. But that part is relatively easy. What is hard to explain is that you don’t care if the money gets used for roads. The second issue is the one I mentioned of open-ended fiscal commitments against the need to have budgets that actually have a bottom line. I think you should perhaps document a little better how other cases work that have dealt with both of these issues work. The IDA 13 replenishment might or might not be a good example. But if I were you I’d try to come up with examples of how this HAS worked in other cases, even if they are cases of a national government giving matching grants to communities. Or of donors literally giving money away for an achievement, rather than inputs. Debt for nature swaps? There must be other cases.